Genomics 111 (2019) 1167–1175

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Methods Paper

DNA physical properties outperform sequence compositional information in classifying nucleosome-enriched and -depleted regions

Guoqing Liu[a,*], Guo-Jun Liu[c], Jiu-Xin Tan[b], Hao Lin[b,**]

[a] The School of Life Science and Technology, Inner Mongolia University of Science and Technology, Baotou 014010, China
[b] Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu, China
[c] School of Natural Sciences and Mathematics, Ural Federal University, Ekaterinburg 620000, Russia

ARTICLE INFO

Keywords:
Nucleosome occupancy
Physical descriptor
Force constant
Flexibility

ABSTRACT

The nucleosome is the fundamental structural unit of eukaryotic chromatin and plays an essential role in the epigenetic regulation of cellular processes, such as DNA replication, recombination, and transcription. Hence, it is important to identify nucleosome positions in the genome. Our previous model based on DNA deformation energy, in which a set of DNA physical descriptors was used, performed well in predicting nucleosome dyad positions and occupancy. In this study, we established a machine-learning model for predicting nucleosome occupancy in order to further verify the physical descriptors. Results showed that (1) our model outperformed several other sequence compositional information-based models, indicating a stronger dependence of nucleosome positioning on DNA physical properties; (2) nucleosome-enriched and -depleted regions have distinct features in terms of DNA physical descriptors like sequence-dependent flexibility and equilibrium structure parameters; (3) gene transcription start sites and termination sites can be well characterized with the distribution patterns of the physical descriptors, indicating the regulatory role of DNA physical properties in gene transcription. In addition, we developed a web server for the model, which is freely accessible at http://lin-group.cn/server/iNuc-force/.

1. Introduction

The nucleosome is the basis of high-order structure of eukaryotic chromatin and plays important roles in gene transcription. It consists of a 147-bp core DNA tightly wrapped ~1.7 times around the histone octamer [1, 2]. Nucleosome beads along the chromatin string are condensed to form 30-nm chromatin structure, which can be further converted into a high-order structure [3]. Nucleosome-directed regulation has been shown to be much more fundamental, ubiquitous and fine-scaled in various processes by modulating the accessibility of genomic sequences to proteins. Nucleosome depletion at the upstream of transcription start sites, a typical regulatory phenomenon, could assist the interaction between transcription factors and DNA regulatory sequences [4].

Due to the increasing amount of experimentally determined nucleosome positions [5–11] and the crucial effect of nucleosome positions on various cellular processes [7, 12–15], in the past 30 years, a lot of models [16–32] for predicting nucleosome positions have been presented. A major class of these models are based on nucleosome positioning signals encoded in DNA sequence [7, 16–36], whereas a few models considered non-sequence factors [35, 36] such as chromatin remodelers and epigenetic marks. The former class (sequence-based models) consists of two kinds of models: bioinformatics models [16–25] and physical models [26–34]. Because these models were constructed based on different principles and different benchmark datasets, they exhibited different prediction capability [37]. The advantages and limitations of the different class of models have been discussed [27] and are not described here.

We initially presented a DNA deformation energy-based model for predicting nucleosome occupancy [38], and then developed an improved model by using a correct form of shearing energy term and re-estimated parameters [30]. Using our computational model, we obtained distribution patterns of nucleosome occupancy around both ends of pseudogenes in the human genome [39], which are in agreement with the experimental results. Analyses based on the model also indicated that processed pseudogenes show a decreased ability to form nucleosomes during evolution [40]. Possible applications of the model in the prediction of nucleosome free energy and nucleosome sliding

* Correspondence to: Guoqing Liu, The School of Life Science and Technology, Inner Mongolia University of Science and Technology, Baotou 014010, China.
** Correspondence to: Hao Lin, Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu, China.
E-mail addresses: gqliu1010@163.com (G. Liu), hlin@uestc.edu.cn (H. Lin).

https://doi.org/10.1016/j.ygeno.2018.07.013
Received 5 March 2018; Received in revised form 7 July 2018; Accepted 15 July 2018
Available online 26 July 2018

were discussed [41]. More importantly, the model was successfully applied to the prediction of nucleosome dyad positions and occupancy [30]. Particularly, bending energy performed well in the dyad position prediction [30]. The successful application of the model could be ascribed to two reasons. Firstly, the established deformation energy model well captured the determinant of nucleosome positioning. Secondly, the pre-estimated parameters for DNA in the model are reliable in nucleosome prediction. Thus, in this study, we still addressed the prediction of nucleosomal regions in *Saccharomyces cerevisiae* (*S. cerevisiae*) using the physical model in order to demonstrate the reliability of the physical descriptors. Furthermore, we discussed the specific relationship between the physical descriptors and nucleosome positioning in detail. We also characterized transcription start sites (TSS) and termination sites (TTS) in terms of the DNA physical descriptors.

2. Materials and methods

2.1. Materials

The complete genome of *S. cerevisiae* (sacCer1 version) was retrieved from UCSC (http://genome.ucsc.edu/). The experimental data of normalized-nucleosome occupancy in vivo across the genome (sacCer1 version) of *S. cerevisiae* was taken from Kaplan et al. [7]. Nucleosome-enriched and -depleted regions were defined as described in Liu et al. [30]. To be specific, according to in vivo nucleosome map [7], if a maximal consecutive regions with the length of > 50 bp has the nucleosome occupancy of > 0.75 or < -0.75 at every site, the region was defined as nucleosome-enriched regions or nucleosome-depleted regions. Based on this criterion, we obtained a benchmark dataset consisting of 16,683 nucleosome-enriched regions and 11,055 nucleosome-depleted regions.

TSS and TTS in yeast were obtained from high-resolution transcription map [42] according to the following procedure. Firstly, we selected poly(A) RNA hybridization-based transcription segments, for which not only there is a clear drop in the hybridization signal (z3, z52 in ref. [42]) at both boundaries, but also at least one of complete coding regions of experimentally verified or uncharacterized genes is contained in their genomic ranges; Secondly, the selected segments that have any overlap with a gene located on the opposite strand were excluded to avoid possible impact of closely located nucleosome free regions around UTR ends which are located on the opposite strand. Thirdly, transcription segments represented by duplicated probes were excluded to filter data with a higher mapping accuracy. According to this way, we finally obtained 1868 segments with mapped 5′ UTR and 3′ UTR, and their start sites and termination sites were defined as TSS and TTS, respectively, of complete transcripts. Genomic sequences around TSS and TTS were retrieved by using their chromosomal coordinates from the complete genome of *S. cerevisiae* (S288C strain).

Genomic coordinates (sacCer3-based) for the precisely identified $-1/+1$ nucleosomes and nucleosome free regions (NFR) in *S. cerevisiae* and corresponding version of the genome (sacCer3 version) were derived, respectively, from the literature [43] and UCSC.

2.2. DNA physical descriptors

DNA physical properties are very useful for studying both the gene structure and chromatin structure [19, 26, 44–46]. DNA physical descriptors used in this study are described below.

According to the Cambridge Convention [47], each base pair in DNA double helix is viewed as a rigid board, and its position relative to its neighbor is specified by six degrees of freedom, such as roll, tilt, twist, slide, shift and rise. Two major kinds of deformations, bending and shear, in nucleosomal DNA were formulated in our previous model [30]. DNA bending is derived largely from roll and tilt. DNA shear is derived from slide and shift. Helical twist determines the phase of each dinucleotide with respect to dyad axis of a nucleosome, and may also play an important role in nucleosome positioning. Therefore, nine physical descriptors, including five equilibrium values (roll, tilt, slide, shift and twist) and four force constants corresponding to roll, tilt, slide and shift, were used to characterize sample sequences. These parameters could be taken from our previous study [30] and listed in Table S1.

One may notice that rise and force constants associated with rise and twist were absent in Table S1. Thus, to construct a complete set of physical descriptors, we re-estimated equilibrium structure parameters and force constants by using a published method [33, 48] based on experimentally determined structures of protein-DNA complexes and free DNA molecules [49]. The resulting two sets of parameters (Table S2, Table S3) were also used for sequence characterization and prediction to make a comparison.

2.3. Quadratic discriminant classifier

We adopted a quadratic discriminant function based on Mahalanobis distance to perform prediction [50]. Mahalanobis distance is based on the correlations among variables and is scale-invariant [50, 51]. It is a useful measure of the difference between an unknown sample set and a known one and has been widely used in classification problems [52–54].

In two-class prediction, there are two training datasets (positive and negative set). Each training dataset is supposed to be represented by N vectors of n dimension, in which the j-th vector is denoted as $X_j^u = [x_{j1}, x_{j2}, \ldots, x_{jn}]^T$ ($j = 1, 2, \ldots, N$; $u = 1,2$). The superscript u in the expression denotes the class of data, that is, $u = 1$ corresponds to positive set and $u = 2$ to negative. The mean vector averaged over the dataset is denoted as $\overline{X^u} = [\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_n]$, where $\overline{x}_i = \sum_{j=1}^{N} x_{ji}/N$, $i = 1, 2, \ldots, n$. The covariance matrix of the u-th training dataset is denoted as

$$C_u = \begin{bmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & \ddots & \vdots \\ c_{n1} & \cdots & c_{nn} \end{bmatrix} \quad (1)$$

where $c_{ji} = \sum_{s=1}^{N} (x_{sj} - \overline{x}_j)(x_{si} - \overline{x}_i)/(N-1)$, $c_{ji} = c_{ij}$.

In the same state space of training datasets, an individual in the test set is represented by a vector of n dimensions: $Y = [y_1, y_2, \ldots, y_n]^T$, and the quadratic discriminant function that determines the class of the individual is given by [54].

$$\xi = \log_2 \frac{N_1}{N_2} - \frac{\delta_1 - \delta_2}{2} - \frac{1}{2}\log_2 \frac{|C_1|}{|C_2|} \quad (2)$$

where N_1 and N_2 respectively, represent the sample sizes of positive and negative training sets,. $\delta_1 = (Y - \overline{X^u})^T (C_u)^{-1} (Y - \overline{X^u})$ is the squared Mahalanobis distance between Y and $\overline{X^u}$, and $|C_u|$ is the determinant of covariance matrix C_u. Let ξ_0 be the optimal threshold of ξ for discriminating two test datasets. Generally, the ξ_0 is around zero for equal-sized positive and negative training sets, and deviates from zero for different-sized samples. In this study, the optimal threshold is determined according to an empirical rule [52]. The test sequences were assigned to positive class if $\xi > \xi_0$, otherwise to negative class.

2.4. Predictive model

Nucleosome-enriched sequences and nucleosome-depleted sequences are respectively labeled as “positive” and “negative”. We carried out two-class prediction by using the quadratic discriminant classifier. The prediction involves two steps. Firstly, each sequence in both positive and negative datasets is represented by a feature vector composed of physical descriptors (Table S1-S3). Secondly, the class of a test sequence is determined by the quadratic discriminant function using the feature vector as input. The 5-fold cross-validation was used to examine the prediction performance. Four indices, namely Sensitivity (S_n), Specificity (S_p), Total accuracy (TA) and Mathew's Correlation coefficient (MCC), were extensively used to quantify the prediction

performance and respectively defined as follows:

$$S_n = TP/(TP + FN) \quad (3)$$

$$S_p = TN/(TN + FP) \quad (4)$$

$$TA = (TP + TN)/(TP + FP + TN + FN) \quad (5)$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TN + FN)(TP + FN)(TN + FP)}} \quad (6)$$

Here, *TP*, *FN*, *TN* and *FP* denote the numbers of True Positive, False Negative, True Negative and False Positive, respectively.

To describe the performance of prediction models across the entire range of ξ_0 of quadratic discriminant, the receiver operating characteristic (ROC) curves were also provided. The quality of the proposed method can be objectively evaluated by measuring the area under the receiver operating characteristic curve (AUC).

3. Results and discussion

3.1. Prediction of nucleosome occupancy

The 5-fold cross-validation was used to evaluate the performance of quadratic discriminant classifier for discriminating nucleosome-enriched regions from nucleosome-depleted regions. We initially investigated the prediction performance of each of nine physical descriptors listed in Table S1. Results in Table 1 showed that all the physical descriptors except tilt and shift have the prediction ability with > 80% accuracy. The roll descriptor could produce the maximum accuracy among the nine descriptors. Subsequently, we investigated the performance by using all the physical descriptor as inputs and achieved an overall accuracy of 91.5% (Table 1). This accuracy is almost equal to that of the roll-based prediction, indicating that roll can alone serve as a best indicator of chromatin states (nucleosome-enriched or nucleosome-depleted). The high prediction accuracy of roll can be explained by the large contribution of the roll parameter to DNA bending [55], which plays an important role in nucleosome positioning. The force constant related to shift also performs well in the prediction, which is probably due to its contribution to DNA shearing energy, which has a close relationship with the nucleosome occupancy [30].

To further illustrate the influence of descriptors on nucleosome positioning, Fig. 1 was drawn to show the distributions of nine physical descriptors between nucleosome-enriched regions and nucleosome-depleted regions. From the figure, we noticed that nucleosome-enriched regions prefer to use the dinucleotides with higher roll, higher slide, lower twist, higher force constant with respect to roll, lower force constants with respect to tilt, slide and shift. Analysis of variance (ANOVA) also confirmed the significant differences (*P*-values < .001) in the physical descriptors between nucleosome-enriched and -depleted regions, except tilt and shift (*P*-values = .72 and 0.96). While we initially speculate that the exceptions for tilt and shift were caused by the assignment of zero for several dinucleotides [30]. To further examine the conclusion, we made predictions by using tilt and shift values (shown in Table S2) estimated by a commonly used method [33]. However, poor predictions (Table 2) were still produced, indicating that tilt and shift values were real poor predictors.

Table 1
The performance of different models in discriminating nucleosome-enriched regions from nucleosome-depleted regions.

Method	Feature	S_n(%)	S_p(%)	*TA*(%)	*MCC*
IDQD[a]	4-mer	84.5	84.7	84.6	0.685
QD[b]	All descriptors	91.7	91.2	91.5	0.825
	f(shift)	91.4	89.8	90.8	0.809
	f(slide)	89.3	86.5	88.2	0.755
	f(roll)	85.6	79.4	83.1	0.649
	f(tilt)	88.9	86.3	87.9	0.749
	shift	61.7	42.0	53.8	0.037
	slide	84.8	79.9	82.8	0.645
	roll	92.0	90.2	91.3	0.820
	tilt	66.2	44.4	57.5	0.107
	twist	87.5	84.5	86.3	0.718
iNuc-PhysChem[c]		100	94.8	97.9	2

[a] Prediction by IDQD using 4-mer frequencies as features [23].
[b] Prediction in this study.
[c] Prediction by iNuc-PhysChem, which used 884 position-specific physicochemical features [19].

3.2. Comparison with sequence-based models

It is necessary to compare the model based on the DNA physical descriptors with other sequence-based models. Hence, we examined the prediction performance of IDQD model [23], which was based on 4-mer oligo-nucleotide frequencies in sequences. The results were also recorded in Table 1. Comparison showed that the prediction accuracy of the physical descriptor-based model increased by 7% when comparing with IDQD model, demonstrating that DNA physical properties used in this study were a better choice of features in discriminating nucleosome-enriched regions from nucleosome-depleted regions. We also investigated the prediction performance of a model called iNuc-PhysChem [19] on our data. Table 1 showed that the model could produce the best prediction among all models listed in Table 1. However, we noticed that the iNuc-PhysChem used 884 position-specific physicochemical features to train the model. The model proposed in this paper used only twelve physical descriptors. Thus, it is not surprising that the iNuc-PhysChem outperforms our proposed model because a high dimensional features contain more information for nucleosome positioning. However, high-dimension features maybe bring out more redundant information, reduce the robust of the model. Our proposed model could not only produce an overall accuracy of 91.5% which was high enough for nucleosome prediction, but also provide a robust prediction because fewer features were used. Thus, we suggested using the model proposed in the paper to perform nucleosome discrimination.

We also made a comparison with other published models [7, 23, 24] by calculating areas under ROC curves (AUC) using the same data. As shown in Fig. 2, our model has the best performance with the AUC of 0.967. Note that in the predictions, we simply used 147-bp (or 130-bp) sequence segments as the input, meaning that possible effects of flanking genomic regions of the sequences and steric exclusion were not considered.

3.3. Further discussion about physical descriptors

A complete set of physical parameters (Table S2) re-estimated from the structures of protein-DNA complexes using a traditional method [33] produced prediction results (Table 2) similar to those presented in Table 1. In the estimation of the complete set of physical parameters, dinucleotide steps were counted from both strands of DNA helix, whereas the previous version of parameters (Table S1) were estimated from one strand of DNA helix. Similar results further suggest the reliability of the estimated parameters. Overall, the two sets of force constants and several equilibrium parameters like roll and twist still remain the high prediction ability.

The physical properties of naked DNA might differ from that of the same DNA molecules bound to proteins because of the physicochemical interaction between the DNA and proteins. For instance, a highly flexible free DNA fragment may show a reduced flexibility in the form of protein-DNA complex if their binding energy (free energy) is low and the formed protein-DNA complex is structurally stiff. To test this, we re-estimated the physical descriptors from free DNA samples (Table S3) and then used the set of re-estimated descriptors to classify nucleosome-enriched/depleted sequences. The correlation between the two sets of physical descriptors was analyzed and showed in Fig. 3. From the

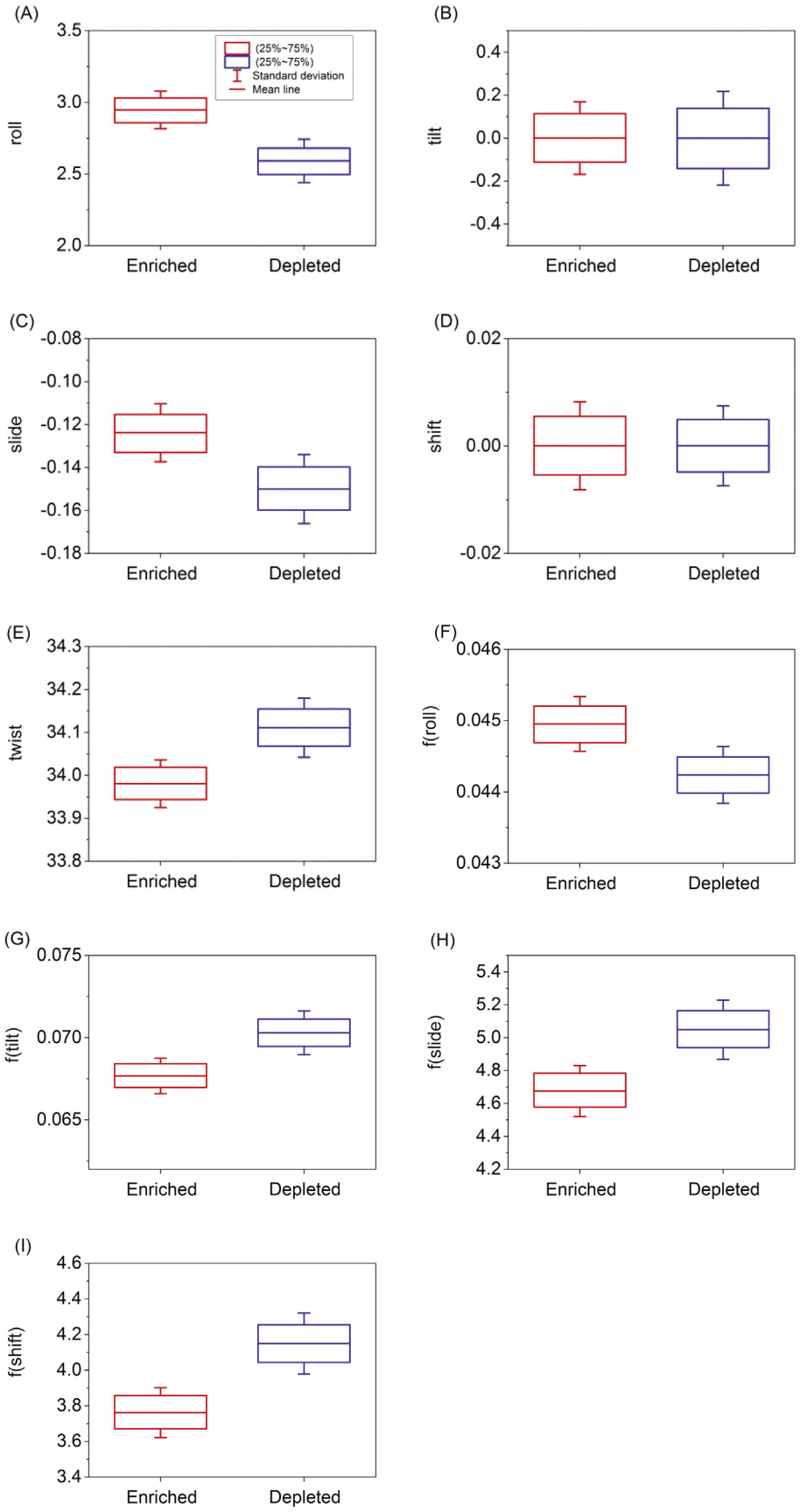

Fig. 1. The difference in physical descriptors between nucleosome-enriched regions and nucleosome-depleted regions. The results of analysis of variance (ANOVA) showed that there are the significant differences (*P*-values < .001) in the physical descriptors between nucleosome-enriched and -depleted regions, except tilt and shift (*P*-values = .72 and 0.96).

Table 2
The performance of physical descriptors derived from structures of protein-DNA complexes in discriminating nucleosome-enriched regions from nucleosome-depleted regions.

Feature	S_n(%)	S_p(%)	TA(%)	MCC
All descriptors	91.6	90.6	91.2	0.817
f(shift)	90.7	88.4	89.8	0.788
f(slide)	89.2	86.5	88.1	0.754
f(rise)	86.1	82.0	84.5	0.678
f(tilt)	88.0	84.7	86.7	0.724
f(roll)	87.2	81.2	84.8	0.684
f(twist)	89.3	84.9	87.5	0.740
Shift	62.1	42.1	54.2	0.043
Slide	84.8	80.0	82.9	0.645
Rise	85.5	82.9	84.4	0.679
Tilt	66.2	44.5	57.5	0.107
Roll	92.0	90.2	91.3	0.819
Twist	87.6	84.6	86.4	0.718

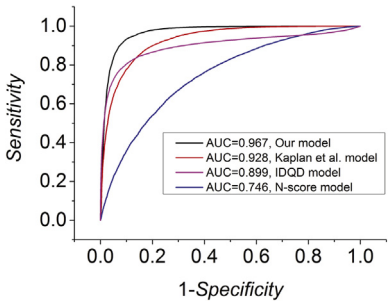

Fig. 2. ROC curve-based comparison of performance between different models in classifying nucleosome-enriched and -depleted regions. Kaplan et al. model, IDQD model, and N-score model refer to those described in the literature [7, 23, 24]. As required by the N-score algorithm [24], only the central 130-bp regions of sequences to be predicted (all of the nucleosome-enriched and -depleted regions described in Materials) are used as input in the corresponding predictions, while in the other predictions central 147-bp regions are used. Probability values calculated in the Kaplan et al. model are used to indicate nucleosome-forming capacity. For N-score and Kaplan et al. models, the data for ROC curves are obtained by adjusting the threshold value of calculated scores, which are used to discriminate nucleosome-enriched regions from nucleosome-depleted regions. For IDQD and present models, ROC curves averaged over 5-fold cross validation results are shown.

figure, we found that the descriptors from naked DNA sequences are strongly correlated with those from protein-DNA complexes, except for rise (green box marked). Moreover, the two sets of force constants have a highly-correlated cluster, suggesting that dinucleotides in free DNA often remain their flexibility level when they are bound to proteins. The shift and tilt have little correlation with other descriptors, and slide and roll are negatively correlated with various force constants. Although the high correlation between the two sets of descriptors was observed, we cannot neglect the difference in the prediction accuracy for some parameters. For example, the prediction accuracies of roll, twist and f (tilt) derived from free DNA structures decreased by 12%, 7% and 7%, respectively (Table 2 and Table 3). This suggests that even a mild alteration of force constants particularly for dinucleotides that play a key role in nucleosome positioning may largely affect the prediction accuracy.

3.4. Web server guide

For the convenience of users, we developed a web server for our prediction model, which can be freely accessible at http://lin-group.cn/server/iNuc-force/. Users can either type or paste their Fasta formatted sequences in the input box (Fig. 4) and then click "Submit" button to predict whether they are nucleosome-enriched sequences or nucleosome-depleted sequences.

Some points regarding the web server are worth noting: (i) The prediction of the server was based on the model trained on the benchmark dataset (see "Materials") which can be freely downloaded from the "Download" menu of the web server; (ii) Given the poor ability of two parameters (shift and tilt) in discriminating nucleosome-enriched regions from the nucleosome-depleted regions, only the remaining 10 descriptors derived from protein-DNA complexes (Table S2) were used to define the feature vector of samples; (iii) The server carries out prediction using a pre-defined 147-bp window with a sliding step of 1 bp along the submitted sequence, and therefore all sequences to be predicted should not be shorter than 147 bp; (iv) The maximum number of submitted sequences and the maximum length of each sequence allowed in the prediction are 50 and 1000 bp, respectively. For more details of the web server, see the web server page (http://lin-group.cn/server/iNuc-force/).

3.5. Distribution of physical descriptors around TSS and TTS

It is important to have an intuitive sense about how the physical descriptors are distributed around TSS and TTS where nucleosomes are characteristically depleted. We show that TSS is characterized by a narrow valley (about 15-bp region) in force constants associated with twist, tilt, slide, rise and shift and two extremely stiff regions surrounded the valley (Fig. 5A, B,). However, the force constant associated with roll displays an opposite behavior that a peak at TSS is surrounded by highly flexible regions (red curve in Fig. 5B). In terms of the distribution of equilibrium parameters, we found that TSS regions have high values of shift, slide, rise, tilt and roll, surrounded by low values (Fig. 5C, D). In contrast, a low valley of twist is observed at TSS, which is surrounded by two peaks (Fig. 5D). For genomic regions around TTS, the detected distribution patterns of all force constants and equilibrium parameters are similar to those for TSS (Fig. 6).

It is worth noting that nucleosomes are selectively depleted at the immediate upstream of TSS [6–8], which are characterized by high force constants regarding twist, tilt, slide, rise and shift, and low force constants regarding roll. In addition, the upstream region of TSS is also characterized by the reduced roll, tilt, slide, rise and shift, and the elevated twist (Fig. 5). These patterns are consistent with the results for experimental data-based nucleosome-enriched and -depleted regions (Fig. 1). These patterns also hold for an extensively identified nucleosome-depleted region at the 3′ end of genes (Fig. 6).

The distribution patterns of physical descriptors estimated from free DNA samples around TSS and TTS were plotted respectively in Fig. S1 and Fig. S2. Similar patterns for free DNA-derived descriptors to those for protein-DNA complexes-derived descriptors were observed (Figs. 5,6).

A previous study demonstrated a clear correlation between molecular dynamics-based physical properties of naked DNA and nucleosome positioning in yeast [56]. Overall, the local structure of DNA around regulatory regions was unusually flexible and displayed a unique pattern of nucleosome positioning. More specifically, the force constants regarding twist and shift were found to be elevated at both the upstream of TSS and downstream of TTS, whereas the others showed a valley at the regions [56]. This observation differs a lot from our results described above, highlighting the importance of correct estimation of force constants for individual dinucleotides. To make a direct comparison with our results, we obtained distribution patterns (Fig. S3) of their force constants [57] around the TSS and TTS data used in this study. Much larger variations in their force constants along the sequences were observed. Some differences and even contradiction exist between the results (Fig. S3) and those based on force constants used in this study (Figs. 5, 6). In an earlier study [57], TSS regions in human genome were reported to have high force constants regarding rise, tilt and roll and low force constants regarding slide, shift and twist [57], which also differ from our results.

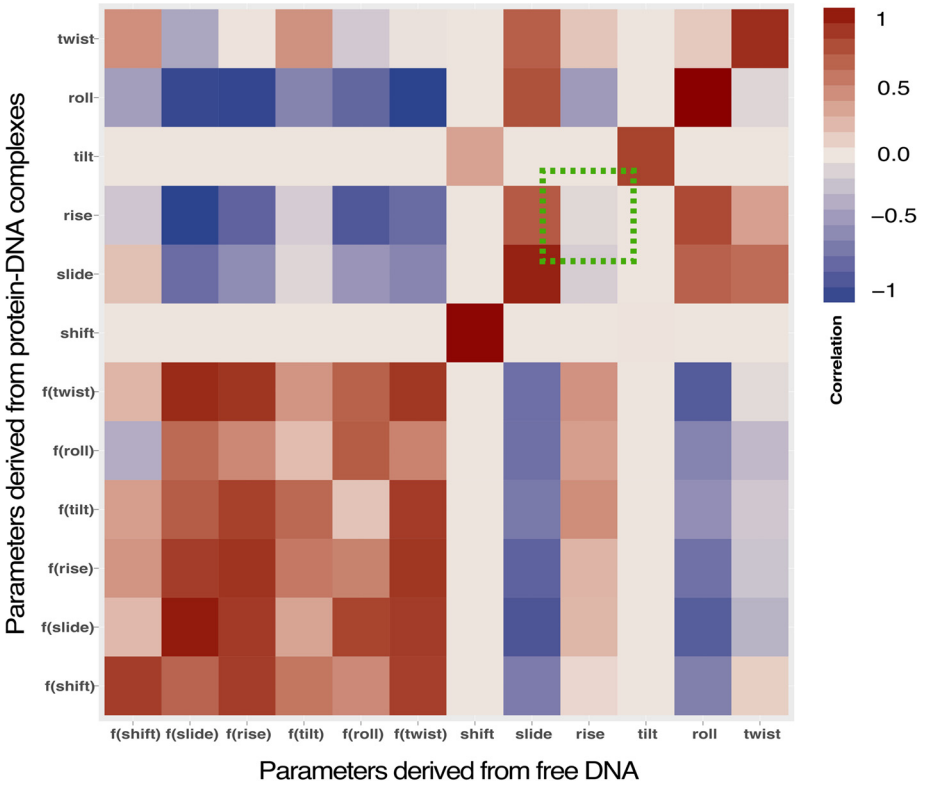

Fig. 3. Spearman correlations between the two sets of physical parameters listed in Table S2 and Table S3. The parameters derived from naked DNA sequences are strongly correlated with those from protein-DNA complexes, except for rise (green box marked). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
The performance of physical descriptors derived from free DNA structures in discriminating nucleosome-enriched regions from nucleosome-depleted regions.

Feature	S_n(%)	S_p(%)	TA(%)	MCC
All descriptors	92.5	90.2	91.6	0.825
f(shift)	89.2	86.5	88.1	0.754
f(slide)	90.1	88.4	89.4	0.781
f(rise)	90.5	88.8	89.9	0.790
f(tilt)	81.6	75.8	79.3	0.571
f(roll)	83.5	78.2	81.4	0.615
f(twist)	91.1	88.5	90.1	0.793
Shift	62.8	44.2	55.4	0.070
Slide	86.9	83.9	85.7	0.704
Rise	81.7	76.9	79.8	0.582
Tilt	66.7	43.6	57.5	0.104
Roll	81.5	75.7	79.2	0.569
Twist	81.1	75.3	78.8	0.561

Despite the inconsistency, our results support the following points. Firstly, regulatory regions in genomes are characterized with unusual levels of physical properties. Secondly, different regulatory regions may have distinct patterns of physical properties although these properties, in some cases, jointly contribute to a single physical factor like deformation energy. Thirdly, physical properties at particular regulatory regions may differ among species. Fourthly, there is no consistent pattern between different force constants for a particular type of regulatory regions. For example, high flexibility at the upstream of TSS is required for some parameters like roll, whereas the rigidity is required for other parameters. This indicates that it is too simple to use general concepts like 'general flexibility' to describe the diverse physical properties. The detailed and unique characteristics regarding different physical properties at regulatory regions are important to understand transcriptional regulation. It should be emphasized that all the phenomena discussed above are a joint effect between physical property-based regulation probably through modulating chromatin structure and

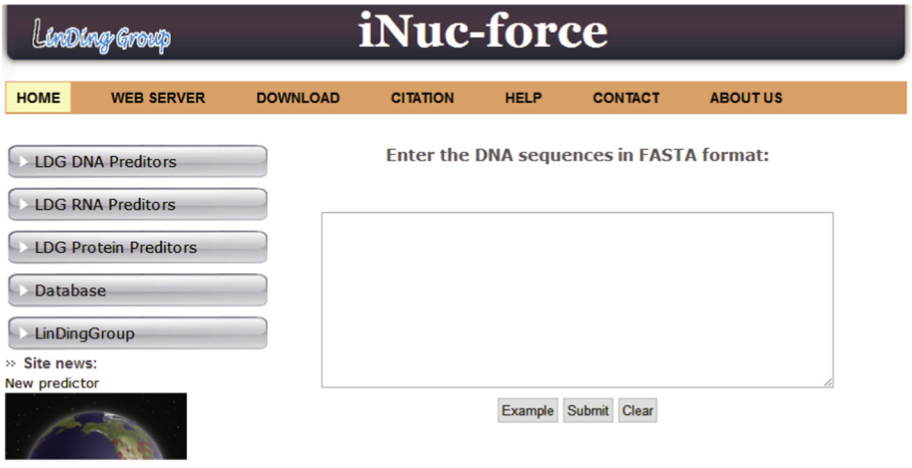

Fig. 4. A screenshot to show the page of the iNuc-force web-server. It is accessible at http://lin-group.cn/server/iNuc-force/predictor.php. Only fasta formatted sequences are allowed to submit, and the server predicts if the submitted sequence is a nucleosome-enriched sequence or a nucleosome-depleted sequences prediction using a pre-defined 147-bp window.

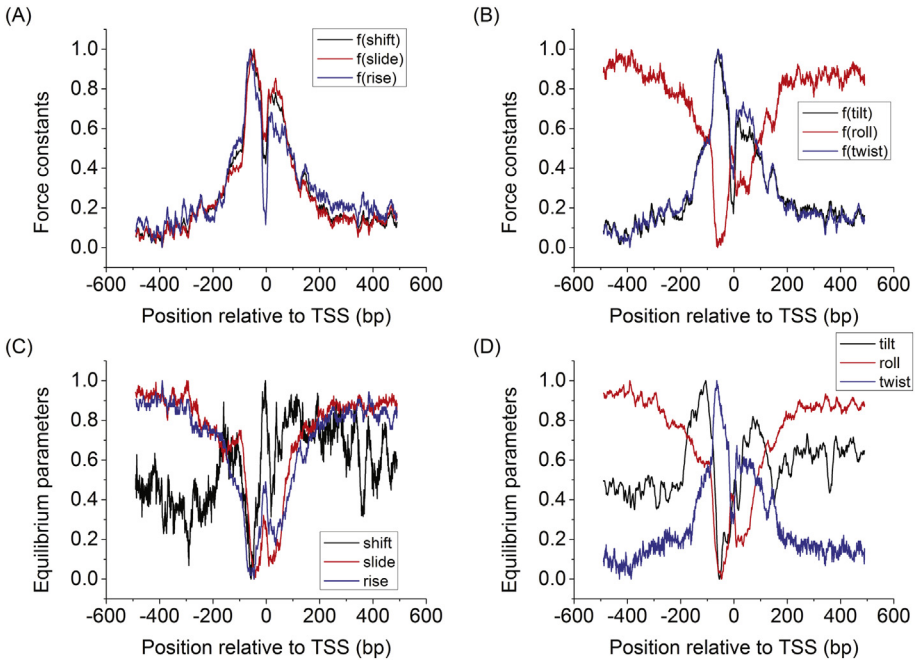

Fig. 5. Distribution of physical descriptors derived from protein-DNA complexes over 1868 transcriptional start sites (TSS) defined in this study (see Materials for details). (A) force constants related to three translational parameters (shift, slide and rise) of DNA structure; (B) force constants related to three rotational parameters (tilt, roll and twist); (C) three translational parameters (shift, slide and rise); (D) three rotational parameters (tilt, roll and twist).

sequence specific readout mechanism in protein-DNA interaction.

The distribution pattern of the physical descriptors around TSS and TTS depends strongly on the nucleosome distribution at the regions, and different yeast genes may have NFRs and + 1/−1 nucleosomes located at different locations relative to the TSS. In order to see how the physical descriptors distribute around the NFR surrounded by −1/+1 nucleosomes, we conducted an analysis on the precisely identified NFRs and − 1/+1 nucleosomes [43]. As compared with the regions where −1/+1 nucleosomes are positioned, the NFR is characterized with larger force constants related to shift, slide, rise, tilt and twist, but with small force constants related to roll (Fig. 7, A-B). This is in consistent with the results shown in Fig. 1 and Fig. 5. Regarding the equilibrium parameters, the NFR preferentially uses dinucleotides with low values of rise and roll, but with high values of twist (Fig. 7, C-D). This is also consistent with our previous results. However, a high peak of slide is detected in NFR, which is inconsistent with the previous results (Fig. 1 and Fig. 5) and needs further investigation. It is interesting that the shift and tilt exhibit a decrease in 5′ to 3′ direction within the NFR (Fig. 7, C-D).

Taken together, our statistical analysis implicates that the nucleosome depletion at TSS and TTS are regulated at least partially by DNA physical properties. Because nucleosome positioning plays important roles in various cellular processes [58] and is coupled to many other epigenetic modifications such as histone modifications [36] and DNA methylation/demethylation [59, 60], DNA physical properties such as DNA stiffness analyzed in our study may provide a new insight into the

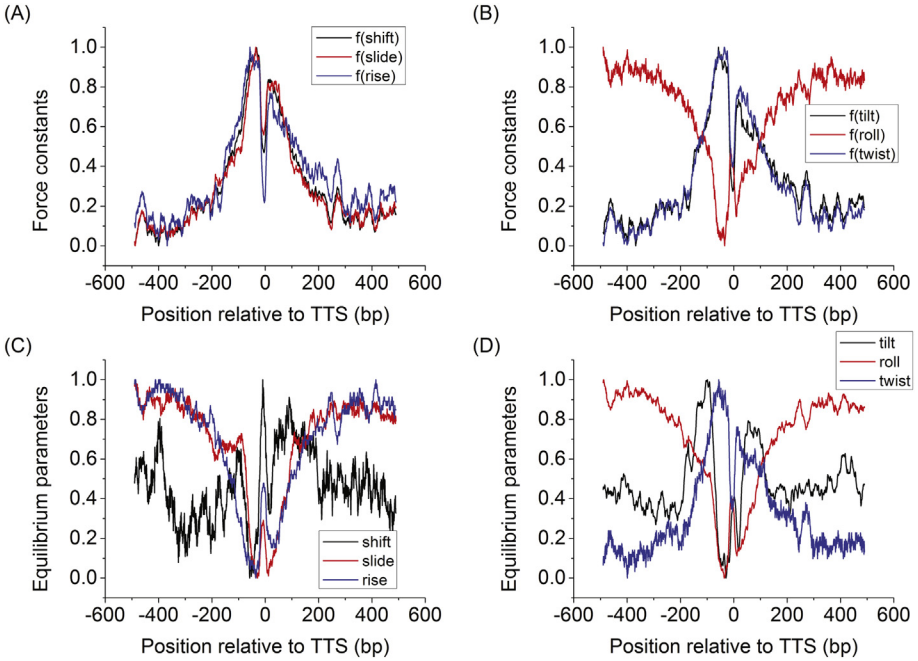

Fig. 6. Distribution of physical descriptors derived from protein-DNA complexes over 1868 transcriptional termination sites (TTS). Sub legends for panels A-D are the same as in Fig. 5.

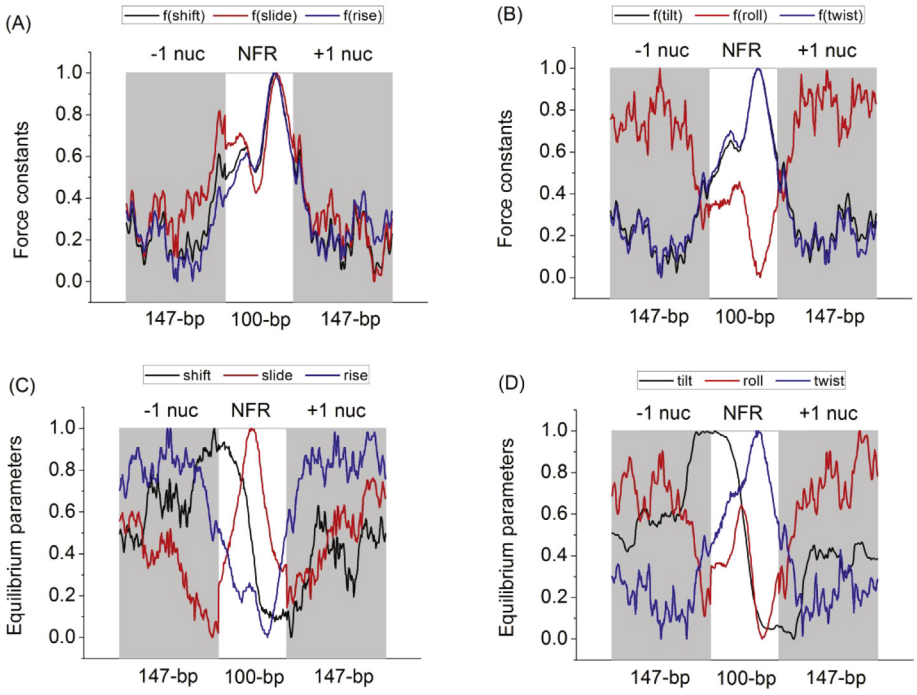

Fig. 7. Distribution of physical descriptors derived from protein-DNA complexes at the positions of nucleosome free regions (NFR) and − 1/+1 nucleosomes (grey shaded area) surrounding the NFR, which were identified in the literature [67]. Three types of the sequences, the numbers of which are all 5542, are aligned, respectively, at their central positions, and then the average distribution pattern of each physical descriptor is obtained for the sequences. Finally, sliding averages are calculated using a window of 20 bp to smooth the distribution curve. The lengths of the three types of sequences analyzed are indicated in the Figure. Sub legends for panels A-D are the same as in Fig. 5.

understanding of eukaryotic chromatins and identification of nucleosome positions.

4. Conclusions

In this study, we demonstrated that physical properties of DNA sequences are the strong indicator of the level of nucleosome occupancy in budding yeast and outperform other features like oligo-nucleotide frequency and periodicity of dinucleotide distribution. We found that physical descriptors like roll and force constants of dinucleotides shift could predict nucleosome occupancy successfully. In addition, we found that TSS and TTS regions in budding yeast could be well characterized by clear distribution patterns of physical properties. These physical descriptors are also likely to be a mark encoded in genomic sequences to mediate nucleosome positioning and regulate gene transcription. Based on our prediction model, a web server was developed and can be freely accessible at http://lin-group.cn/server/iNuc-force/. This predictor will play at least a complementary role to the existing approaches in predicting nucleosome-enriched and nucleosome-depleted regions.

Acknowledgements

This work was supported by grants from the National Natural Science Foundation (31660322, 61102162, 61772119) and Science Foundation for Excellent Youth Scholars of Inner Mongolia University of Science and Technology (2016YQL06).

Author contributions

GL and HL designed the study, carried out the calculation and wrote the manuscript. GJL and JXT participated in the data analysis. HL and GL developed the web-server.

Additional information

Competing financial interests: The authors declare no competing financial interests.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2018.07.013.

References

[1] T.J. Richmond, C.A. Davey, The structure of DNA in the nucleosome core, Nature 423 (2003) 145–150.
[2] R.D. Kornberg, Y. Lorch, Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome, Cell 98 (1999) 285–294.
[3] K. Maeshima, R. Imai, S. Tamura, T. Nozaki, Chromatin as dynamic 10-nm fibers, Chromosoma 123 (2014) 225–237.
[4] C.K. Lee, Y. Shibata, B. Rao, B.D. Strahl, J.D. Lieb, Evidence for nucleosome depletion at active regulatory regions genome-wide, Nat. Genet. 36 (2004) 900–905.
[5] G.C. Yuan, Y.J. Liu, M.F. Dion, M.D. Slack, L.F. Wu, S.J. Altschuler, et al., Genome-scale identification of nucleosome positions in S. Cerevisiae, Science 309 (2005) 626–630.
[6] W. Lee, D. Tillo, N. Bray, R.H. Morse, R.W. Davis, T.R. Hughes, et al., A high-resolution atlas of nucleosome occupancy in yeast, Nat. Genet. 39 (2007) 1235–1244.
[7] N. Kaplan, I.K. Moore, Y. Fondufe-Mittendorf, A.J. Gossett, D. Tillo, Y. Field, et al., The DNA-encoded nucleosome organization of a eukaryotic genome, Nature 458 (2009) 362–366.
[8] T.N. Mavrich, I.P. Ioshikhes, B.J. Venters, C. Jiang, L.P. Tomsho, J. Qi, et al., A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome, Genome Res. 18 (2008) 1073–1083.
[9] K. Brogaard, L. Xi, J.P. Wang, J. Widom, A map of nucleosome positions in yeast at base-pair resolution, Nature 486 (2012) 496–501.
[10] A. Valouev, J. Ichikawa, T. Tonthat, J. Stuart, S. Ranade, H. Peckham, et al., A high-resolution, nucleosome position map of C. Elegans reveals a lack of universal sequence-dictated positioning, Genome Res. 18 (2008) 1051–1063.
[11] D.E. Schones, K. Cui, S. Cuddapah, T.Y. Roh, A. Barski, Z. Wang, et al., Dynamic regulation of nucleosome positioning in the human genome, Cell 132 (2008) 887–898.
[12] D.M. MacAlpine, G. Almouzni, Chromatin and DNA replication, Cold Spring Harb. Perspect. Biol. 5 (2013) a010207.
[13] C. Peng, H. Luo, X. Zhang, F. Gao, Recent advances in the genome-wide study of DNA replication origins in yeast, Front. Microbiol. 6 (2015) 117.
[14] T. Yamada, K. Ohta, Initiation of meiotic recombination in chromatin structure, Biochemistry 154 (2013) 107–114.
[15] S. Naftelberg, I.E. Schor, G. Ast, A.R. Kornblihtt, Regulation of alternative splicing through coupling with transcription and chromatin structure, Annu. Rev. Biochem. 84 (2015) 165–198.
[16] H.E. Peckham, R.E. Thurman, Y. Fu, J.A. Stamatoyannopoulos, W.S. Noble, K. Struhl, et al., Nucleosome positioning signals in genomic DNA, Genome Res. 17 (2007) 1170–1177.
[17] Y. Xing, X. Zhao, L. Cai, Prediction of nucleosome occupancy in Saccharomyces cerevisiae using position-correlation scoring function, Genomics 98 (2011)

359–366.
[18] S.H. Guo, E.Z. Deng, L.Q. Xu, H. Ding, H. Lin, W. Chen, et al., iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition, Bioinformatics 30 (2014) 1522–1529.
[19] W. Chen, H. Lin, P.M. Feng, C. Ding, Y.C. Zuo, K.C. Chou, iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via physicochemical properties, PLoS One 7 (2012) e47843.
[20] I. Gabdank, D. Barash, E.N. Trifonov, Single-base resolution nucleosome mapping on DNA sequences, Biomol Struct Dyn 28 (2010) 107–121.
[21] T. van der Heijden, J.J. van Vugt, C. Logie, J. van Noort, Sequence-based prediction of single nucleosome positioning and genome-wide nucleosome occupancy, Proc. Natl. Acad. Sci. 109 (2012) E2514–E2522.
[22] F. Cui, L. Chen, P.R. Loverso, V.B. Zhurkin, Prediction of nucleosome rotational positioning in yeast and human genomes based on sequence-dependent DNA anisotropy, BMC Bioinformatics 15 (2014) 313.
[23] X. Zhao, Z. Pei, J. Liu, S. Qin, L. Cai, Prediction of nucleosome DNA formation potential and nucleosome positioning using increment of diversity combined with quadratic discriminant analysis, Chromosom. Res. 18 (2010) 777–785.
[24] G.C. Yuan, J.S. Liu, Genomic sequence is highly predictive of local nucleosome depletion, PLoS Comput. Biol. 4 (2008) e13.
[25] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thåström, Y. Field, I.K. Moore, et al., A genomic code for nucleosome positioning, Nature 442 (2006) 772–778.
[26] C. Anselmi, G. Bocchinfuso, P. De Santis, M. Savino, A. Scipioni, Dual role of DNA intrinsic curvature and flexibility in determining nucleosome stability, J. Mol. Biol. 286 (1999) 1293–1301.
[27] P. De Santis, S. Morosetti, A. Scipioni, Prediction of nucleosome positioning in genomes. Limits and perspectives of physical and bioinformatic approaches, J. Biomol. Struct. Dyn. 27 (2010) 747–764.
[28] Y.V. Sereda, T.C. Bishop, Evaluation of elastic rod models with long range interactions for predicting nucleosome stability, J. Biomol. Struct. Dyn. 27 (2010) 867–887.
[29] G. Chevereau, L. Palmeira, C. Thermes, A. Arneodo, C. Vaillant, Thermodynamics of intra-genic nucleosome ordering, Phys. Rev. Lett. 103 (2009) 188103.
[30] G. Liu, Y. Xing, H. Zhao, J. Wang, Y. Shang, L. Cai, A deformation energy-based model for predicting nucleosome dyads and occupancy, Sci. Rep. 6 (2016) 24133.
[31] V. Miele, C. Vaillant, D'Aubenton-Carafa Y, Thermes C, Grange T., DNA physical properties determine nucleosome occupancy from yeast to fly, Nucleic Acids Res. 36 (2008) 3746–3756.
[32] M.Y. Tolstorukov, A.V. Colasanti, D.M. McCandlish, W.K. Olson, V.B. Zhurkin, A novel roll-and-slide mechanism of DNA folding in chromatin: implications for nucleosome positioning, J. Mol. Biol. 371 (2007) 725–738.
[33] A.V. Morozov, K. Fortney, D.A. Gaykalova, V.M. Studitsky, J. Widom, E.D. Siggia, Using DNA mechanics to predict in vitro nucleosome positions and formation energies, Nucleic Acids Res. 37 (2009) 4707–4722.
[34] T.C. Bishop, Geometry of the nucleosomal DNA superhelix, Biophys. J. 95 (2008) 1007–1017.
[35] K. Rippe, A. Schrader, P. Riede, R. Strohner, E. Lehmann, G. Längst, DNA sequence- and conformation-directed positioning of nucleosomes by chromatin-remodeling complexes, Proc. Natl. Acad. Sci. U. S. A. 104 (2007) 15635–15640.
[36] Y. Zhang, H. Shin, J.S. Song, Y. Lei, X.S. Liu, Identifying positioned nucleosomes with epigenetic marks in human from ChIP-Seq, BMC Genomics 9 (2008) 537.
[37] H. Liu, R. Zhang, W. Xiong, J. Guan, Z. Zhuang, S. Zhou, A comparative evaluation on prediction methods of nucleosome positioning, Brief. Bioinform. 15 (2014) 1014–1027.
[38] J.Y. Wang, J. Wang, G. Liu, Calculation of nucleosomal DNA deformation energy: its implication for nucleosome positioning, Chromosom. Res. 20 (2012) 889–902.
[39] G. Liu, F. Feng, X. Zhao, L. Cai, Nucleosome organization around pseudogenes in the human genome, Biomed. Res. Int. 821596 (2015).
[40] G. Liu, X. Cui, H. Li, L. Cai, Evolutionary direction of processed pseudogenes, Sci. China Life Sci. 59 (2016) 839–849.
[41] G. Liu, Y. Xing, H. Zhao, L. Cai, J. Wang, The implication of DNA bending energy for nucleosome positioning and sliding, Sci. Rep. 8 (2018) 8853.
[42] L. David, W. Huber, M. Granovskaia, J. Toedling, C.J. Palm, L. Bofkin, et al., A high-resolution map of transcription in the yeast genome, Proc. Natl. Acad. Sci. 103 (2006) 5320–5325.
[43] R.V. Chereji, S. Ramachandran, T.D. Bryson, S. Henikoff, Precise genome-wide mapping of single nucleosomes and linkers in vivo, Genome Biol. 19 (2018) 19.
[44] Y.C. Zuo, Q.Z. Li, The hidden physical codes for modulating the prokaryotic transcription initiation, Physica A 389 (2010) 4217–4223.
[45] Y.C. Zuo, Q.Z. Li, Identification of TATA and TATA-less promoters in plant genomes by integrating diversity measure, GC-skew and DNA geometric flexibility, Genomics 97 (2011) 112–120.
[46] Y.C. Zuo, P. Zhang, L. Liu, T. Li, Y. Peng, G. Li, Q.Z. Li, Sequence-specific flexibility organization of splicing flanking sequence and prediction of splice sites in the human genome, Chromosom. Res. 22 (2014) 321–334.
[47] R.E. Dickerson, Definitions and nomenclature of nucleic acid structure components, Nucleic Acids Res. 17 (1989) 1797–1803.
[48] W.K. Olson, A.A. Gorin, X.J. Lu, L.M. Hock, V.B. Zhurkin, DNA sequence-dependent deformability deduced from protein-DNA crystal complexes, Proc. Natl. Acad. Sci. U. S. A. 95 (1998) 11163–11168.
[49] H.M. Berman, W.K. Olson, D.L. Beveridge, J. Westbrook, A. Gelbin, T. Demeny, et al., The nucleic acid database: a comprehensive relational database of three-dimensional structures of nucleic acids, Biophys. J. 63 (1992) 751–759.
[50] M.Q. Zhang, Identification of protein coding regions in the human genome by quadratic discriminant analysis, Proc. Natl. Acad. Sci. 94 (1997) 565–568.
[51] P.C. Mahalanobis, On the generalised distance in statistics, Proc Natl Inst Sci India 2 (1936) 49–55.
[52] J. Lu, L.F. Luo, L.R. Zhang, W. Chen, Y. Zhang, Increment of diversity with quadratic discriminant analysis-an efficient tool for sequence pattern recognition in bioinformatics, Open Access Bioinformatics 2 (2010) 89–96.
[53] H. Lin, Q.Z. Li, Predicting conotoxin superfamily and family by using pseudo amino acid composition and modified Mahalanobis discriminant, Biochem. Biophys. Res. Commun. 354 (2007) 548–551.
[54] L.R. Zhang, L.F. Luo, Splice site prediction with quadratic discriminant analysis using diversity measure, Nucleic Acids Res. 31 (2003) 6214–6220.
[55] F. Battistini, C.A. Hunter, E.J. Gardiner, M.J. Packer, Structural mechanics of DNA wrapping in the nucleosome, J. Mol. Biol. 396 (2010) 264–279.
[56] Ö. Deniz, O. Flores, F. Battistini, A. Pérez, M. Soler-López, M. Orozco, et al., Physical properties of naked DNA influence nucleosome positioning and correlate with transcription start and termination sites in yeast, BMC Genomics 12 (2011) 489.
[57] J.R. Goñi, A. Pérez, D. Torrents, M. Orozco, Determining promoter location based on DNA structure first-principles calculations, Genome Biol. 8 (2007) R263.
[58] M. Radman-Livaja, O.J. Rando, Nucleosome positioning: how is it established, and why does it matter? Dev. Biol. 339 (2010) 258–266.
[59] Liu D, Li G, Zuo Y. Function determinants of TET proteins: the arrangements of sequence motifs with specific codes, Brief. Bioinform., DOI: https://doi.org/10.1093/bib/bby053.
[60] C. Lövkvist, K. Sneppen, J.O. Haerter, Exploring the link between nucleosome occupancy and DNA methylation, Front. Genet. 8 (2018) 232.